AsiaInfo 亚信科技

AIR 清华大学智能产业研究院 Institute for AI Industry Research, Tsinghua University

Test Report - DeepSeek Empowers the Advanced Evolution of Autonomous Networks

This report integrates the business needs of high - value scenarios in Autonomous Networks and evaluates the technical capabilities of DeepSeek and several typical large models from multiple dimensions such as semantic parsing, intent recognition, and reasoning ability. This is the first evaluation report focusing on the empowerment of large models in the field of Autonomous Networks. It aims to provide scientific basis and technical references for the advanced evolution of Autonomous Networks to L4/L5, and to promote the intelligent transformation and sustainable development of the communication industry.

Authors

AsiaInfo Technologies (China) Ltd.

Zhiqi Li,Xianlei Yang,Lei Zhang,Xiaozhou Ye,Dajiang Du,Zhigang Wang, Guanghui Zhang,Hehe Chen,Xinjian Li,Yang Bai,Yue Zhang,Yan Lu, Duozhi Zhu,Ye Ouyang

Institute for AI Industry Research, Tsinghua University

Yunxin Liu, Xianyuan Zhan, Yuanchun Li, Yuanzhe Li, Weijun Wang, Ce Zhang, Xiangyu Zhu, Jichao Leng

Citation of this evaluation report:

Test Report - DeepSeek Empowers the Advanced Evolution of Autonomous Networks, Ye Ouyang, Yunxin Liu, et al, 2025.2.

Content

Abstract

DeepSeek V3 and R1 have quickly gained prominence in the telecom industry, driven by their high performance, open-source innovation, and cost-efficiency. AsiaInfo's Usights · Advanced Autonomous Network product (AISWare AN Evo[1]) has been full-stack adapted with DeepSeek V3 and R1. To assess the technical performance and application potential of DeepSeek in empowering autonomous networks (AN[2]) for intelligent transformation, this whitepaper examines and research DeepSeek-enabled autonomous networks in the context of AN Evo empowering high-value application by LLM. The test covers multi-dimensions, such as intent recognition, autonomous planning, and reasoning; and results demonstrate that DeepSeek performs excellently in many aspects, while remaining potential for improvement in response speed and efficiency. With further optimization, it is expected that DeepSeek will be a powerful engine for the high-level autonomous network evolution.

Given the rapid evolution of LLM and diverse applications, the conclusions in this paper are specific to the current testing environment. Due to capacity and resource constraints in the team, the analysis may be limited. We welcome feedback from industry professionals for further enhancement.

1. Introduction to DeepSeek

DeepSeek is a Chinese technology company focusing on Artificial General Intelligence (AGI) invested by High-Flyer, aiming to develop advanced large language models (LLM) and related technologies.

The core technology of DeepSeek combines LLM and retrieval engines to augment the knowledge base of LLM through real-time retrieval, addressing the illusion and low time-sensitivity of traditional LLM. Its LLM products include DeepSeek-R1, DeepSeek-V3, and so on. In many benchmark tests, DeepSeek’s multi-dimensional performance is on par with OpenAI's GPT models, and has already surpassed in some areas, but its training cost is only

[1]AISWare AN Evo：AISWare Autonomous Networks Evolution

[2]AN: Autonomous Networks

10% of that of GPT-4. The high cost-effectiveness, cost advantage, and an open-source strategy have driven its rapid commercialization.

DeepSeek is widely applied to various fields such as natural language processing, machine learning, and coding tasks, and is capable of intelligent conversations, accurate translations, creative writing, efficient programming, intelligent puzzle solving, and document reading. Simultaneously, its open-source strategy fosters collaboration and development of a global AI developer community.

Currently, DeepSeek has attracted widespread attention in the AI field with its efficient, open-source LLM and the prospect of its technological development and application is highly expectable.

2. Advancing AN Evolution Towards L4

The goal of autonomous networks is to develop end-to-end automatic and intelligent network operation and maintenance (O&M) capability for the full lifecycle. At present, the autonomous network is undergoing the transformation from traditional L3 to advanced L4, and it is facing several problems, such as a single human-computer interaction (HCI), lack of cognitive Parsing and logical reasoning, insufficient knowledge, and weak generalization. DeepSeek, with its generalized capabilities such as excellent intent Parsing, innovative HCI and augmented specialized knowledge, can facilitate the advanced evolution of autonomous networks with strong technological support.

3. Testing Objective

The main purpose of this test is to analyze the adaptability of DeepSeek in high-value scenarios of autonomous networks, explore its application potential in key scenarios such as network orchestration, network data querying, fault diagnosis, and complaint response, and reduce other vendors' test complexity. This paper aims to provide practical experience for industry applications and promote the technological pervasiveness of the telecommunications industry with overall competitiveness.

4. Product Portfolio of AN Evo

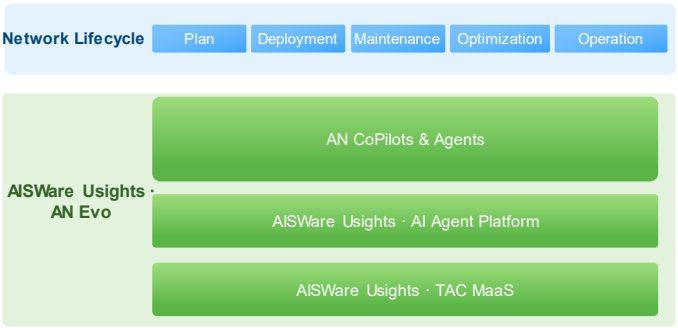


Figure 4-1 Product portfolio of Usights · AISWare AN Evo

AISWare AN Evo is a comprehensive toolkit of CoPilot and AI Agent developed by LLM for the evolution of advanced autonomous networks. The AN CoPilot can assist O&M personnel in network O&M, and the AN Agent is an innovative scenario-oriented design to achieve autonomous decision-making and meet business objectives.

AN Evo employs COTS LLMs as the foundation. It interfaces with DeepSeek to test its performance in autonomous network scenarios.

5. Testing Solutions of DeepSeek Empowering AN

5.1 Objectives

The core objective is to comprehensively assess the technical performance and application potential of DeepSeek in empowering autonomous networks, and to provide a scientific rationale for the advanced autonomous network evolution towards L4/L5.

The specific testing objectives include:

- Test and validate DeepSeek's utility with capabilities in semantic Parsing, analytics and reasoning, multi-round Q&A, intent recognition, autonomous planning, knowledge retrieval, and text generation in an autonomous network application scenario.

- Compare several typical LLM capability pairs to conclude the adaptability of the models in various application scenarios of the current autonomous network.
- Record the response time of LLMs under different tasks and evaluate whether its efficiency meets the actual business requirements.

Through above testing objectives, this solution seeks to reflect the adaptability and advantages of DeepSeek in autonomous networks holistically and objectively, and to provide a reference for its further optimization and practical deployment.

5.2 Testing Environment Setup

5.2.1 Hardware Environment

This test was run with an NVIDIA A800 80GB GPU, a total of 16 GPUs with the following key specifications:

- Core architecture: GA100 with an Ampere-based architecture
- CUDA core counts: 6912
- Graphics memory capacity: 80GB HBM2e
- Graphics memory bandwidth: 1935GB/s.
- Electricity consumption: 250W max
- API: PCIe 4.0 x16

5.2.2 Software Environment

The software configurations of the testing environment:

- CUDA: For support of GPU-accelerated computation
- vLLM: Efficient language model inference engine with multilingual

support

- PyTorch: Deep learning framework for model training and inference
- Flash Attention: For optimized attention mechanism to upgrade the performance of Transformer
- Transformers: Provide a library of pre-trained models with multilingual and multimodal task support

5.2.3 Selection of LLMs

This test selected five representative LLMs covering different scales, architectures and application scenarios for an all-round assessment of performance and adaptability. The selected LLMs are as follows:

- DeepSeek-R1
- DeepSeek-v3
- Three typical LLMs from mainstream open-source models and commercial models: C/D/E (LLM C is the existing model after production debugging)

5.3 General DeepSeek Capabilities

Based on the current practical needs of autonomous network application scenarios, this test mainly examines the following six capabilities: semantic Parsing, analytic and reasoning, multi-round Q&A, intent recognition, autonomous planning, knowledge retrieval, and text generation. The testing objectives, measurements and evaluation indexes of the functional requirements are described as follows.

5.3.1 AN Semantic Parsing

- **Multi-round Conversation Parsing**

1) Testing Objectives

To test the semantic parsing capability of the LLM in multiple rounds of conversation and assess whether it can accurately understand the context and generate appropriate responses.

2) Testing Measurements

Test by dataset from multi-round conversation to calculate the performance of the model on semantic coherence.

3) Evaluation Indexes

Semantic coherence: Whether the model generates a consistent response with the context.

$$Semantic\ coherence = \frac{Generate\ conversation\ counts\ with\ answers\ aligned\ to\ the\ context}{Total\ conversation\ counts} * 100\%$$

Accuracy: Whether the model understands the user's intent accurately and answers correctly.

$$Accuracy = \frac{Correct\ question\ counts\ answered\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Completeness: Whether the model can completely solve the user's problem in multiple conversation rounds.

$$Completeness = \frac{Conversation\ counts\ for\ successful\ problem - solving\ by\ the\ LLM}{Total\ conversation\ counts} * 100\%$$

Note: This test evaluates the model's performance in complex conversation scenarios with a dataset from multi-round conversations, mainly focusing on whether LLM can track the conversation history and generate coherent and sensible responses.

- **Synonym/Near-synonym Identification**

1) Testing Objectives

To test whether the model can identify synonyms or near-synonyms and respond without semantic changes.

2) Testing Measurements

Apply quiz pairs containing synonyms to examine whether the model understands the question correctly after synonym substitution.

3) Evaluation Indexes

Identification rate: The capability of the model to recognize synonyms and near-synonyms

$$Identification\ rate = \frac{Question\ counts\ of\ correct\ identification\ of\ synonyms\ or\ near-synonyms\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Semantic consistency: Whether the model still understands the question correctly after substituting synonyms.

$$Semantic\ consistency\ = \frac{Question\ counts\ of\ correct\ comprehension\ after\ substituting\ synonyms\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Note: This test utilizes quiz pairs containing synonyms to examine the model performance and assess its robustness under semantic changes.

5.3.2 AN Intent Recognition

- **Intent Recognition of Task-orientated Conversations**

1) Testing Objectives

To evaluate the accuracy in recognizing user intent in task-oriented conversations, such as retrieving metadata of network element and troubleshooting.

2) Testing Measurements

Compare the model accuracy in recognizing user intent in each task scenario.

3) Evaluation Indexes

Intent recognition accuracy: Whether the model can correctly recognize user intent.

$$Intent\ recognition\ accuracy\ = \frac{Correct\ intent\ recognition\ by\ the\ LLM}{Total\ testing\ counts} * 100\%$$

Note: This test evaluates the intent recognition capability of LLMs in practical business scenarios (e.g., retrieving metadata of network element and troubleshooting).

- **Entity Extraction and Slot Filling**

1) Testing Objectives

To evaluate the LLM's key information extraction capability (e.g., time, location, network element name, failure number, etc.) from conversations.

2) Testing Measurements

Compare the entities extracted by the model with the standard answers to assess the accuracy.

3) Evaluation Indexes

Extraction accuracy: Whether the model can correctly extract key information (e.g., time, location, network element name, etc.).

$$Extraction\ accuracy\ = \frac{Correct\ extracted\ entity\ counts\ by\ the\ LLM}{Total\ entity\ counts\ in\ standard\ answers} * 100\%$$

Coverage: Whether the model can completely extract all relevant entities.

$$Coverage = \frac{Extracted\ entity\ category\ counts\ by\ the\ LLM}{Total\ entity\ category\ counts\ in\ standard\ answers} * 100\%$$

Note: This test compares the output entities by the model with the standard answers and evaluates its accuracy and recall on entity extraction.

- **Intent-entity Association Identification**

1) Testing Objectives：

To test whether the model can correctly associate to the corresponding entities while recognizing the user's intent.

2) Testing Measurements：

Examine whether the model correctly identifies and associates relevant entities when retrieving specific information.

3) Evaluation Indexes

Association accuracy: Whether the model can correctly associate intent with relevant entities.

$$Association\ accuracy = \frac{Correct\ intent - entity\ association\ times\ by\ the\ LLM}{Total\ test\ times} * 100\%$$

Note: This test examines the LLM's capability to associate intents with entities in complex tasks through specific inquiry scenarios.

5.3.3 AN Reasoning Capability

- **Commonsense Reasoning**

1) Testing Objectives

To test the model capability to deduce reasonable answers based on common sense in non-obvious and implicit information scenarios.

2) Testing Measurements

Ask commonsense questions to assess the model's reasoning performance.

3) Evaluation Indexes

Inference reasonability: Whether the model can infer reasonable answers based on common sense.

$$Inference\ reasonability = \frac{Question\ counts\ to\ inferred\ reasonable\ answers\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Implicit information Parsing: Whether the model can understand non-explicit information.

$$Implicit\ information\ comprehension = \frac{Question\ counts\ to\ correct\ implicit\ information\ comprehension\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Note: This test evaluates the model's reasoning performance by asking commonsense questions, especially in implicit information scenarios.

- **Causal Reasoning**

1) Testing Objectives

To test LLM’s inference capability of causality or chronological order under given presuppositions.

2) Testing Measurements

Evaluate whether the model can infer reasonable processing steps by providing scenarios such as troubleshooting steps.

3) Evaluation Indexes

Causality accuracy: Whether the model can correctly infer causality or chronology.

$$Formula: Causality\ accuracy = \frac{Times\ of\ correct\ inference\ of\ causality\ or\ chronology\ by\ the\ LLM}{Total\ test\ times} * 100\%$$

Note: This test examines the LLM's causality inference capability through scenarios such as troubleshooting steps and verifies its performance in complex tasks.

5.3.4 AN Autonomous Planning

Autonomous planning capability is defined as the capability of an LLM to generate step-by-step solutions based on existing knowledge and input information under a given goal or task. It is of critical importance in autonomous networks, applied in network optimization, troubleshooting, resource scheduling and so on.

1) Testing Objectives：

To evaluate whether the model can generate reasonable optimization plan based on the current state and target requirements in complex network environments. This test examines the functionality of network optimization, troubleshooting, resource scheduling, and service activation.

2) Testing Measurements

Provide a simulation environment or real dataset with relevant scenario services (one or several scenarios of network optimization, troubleshooting, resource scheduling, and service activation, such as traffic load, delay, package loss rate, etc.).

Given an objective of the solution plan (e.g., reduce latency, increase bandwidth utilization, etc.), the model is required to generate specific optimization steps or subsequent solution plan.

Verify whether the output plan by LLM is reasonable and validate the effect through simulation or actual execution.

3) Evaluation Indexes

Plan reasonability: Whether the generated optimization plan by the model meets the actual demand.

$$\text{Plan reasonability} = \frac{\text{Reasonable optimization plan counts generated by the LLM}}{\text{Total test times}} * 100\%$$

Diagnosis accuracy: Whether the model can correctly analyze the root cause of the failure.

$$Diagnosis\ accuracy = \frac{Times\ of\ correct\ failure\ diagnosis\ by\ the\ LLM}{Total\ test\ times} * 100\%$$

Timeliness: Whether the result return speed of the model meets the real-time demand.

$$Timeliness\ = \frac{Times\ of\ result\ return\ within\ a\ defined\ time\ frame\ by\ the\ LLM}{Total\ question\ counts} * 100\%$$

Note: This test evaluates the LLM's autonomous planning capability in complex network environments by testing datasets from network optimization, troubleshooting, and resource scheduling scenarios.

5.3.5 AN Knowledge Retrieval

LLM is capable of extracting relevant information quickly and accurately from massive data.

1) Testing Objectives

To evaluate whether the model is able to retrieve relevant and appropriate information from the existing knowledge base in response to an input question or requirement.

2) Testing Measurements

Provide diversified question sets, covering FAQs (e.g., equipment configuration, troubleshooting), complex questions (e.g., multi-domain co-optimization), and unusual questions (e.g., parameter tuning in specific scenarios).

Verify the accuracy, comprehensiveness, and timeliness of the returned results in a form of Q&A.

3) Evaluation Indexes

Accuracy: Whether the returned results are correct.

$$Accuracy = \frac{Correct\ returned\ results\ counts\ from\ the\ LLM}{Total\ question\ counts} * 100\%$$

Comprehensiveness: Whether the returned results cover all relevant information.

$$Comprehensiveness = \frac{Comprehensive\ returned\ result\ counts\ from\ the\ LLM}{Total\ question\ counts} * 100\%$$

Timeliness: Whether the speed at result returning from the model meets the real-time requirements.

$$Timeliness = \frac{Times\ of\ returned\ results\ from\ the\ LLM\ in\ a\ defined\ time\ frame}{Total\ question\ counts} * 100\%$$

Note: This test evaluates the model's knowledge retrieval capabilities through diversified sets of questions (common, complex, or unusual).

5.3.6 AN Text Generation

LLM can generate coherent, accurate, and contextualized content based on input information. Typical application scenarios of this capability in autonomous networks include operation manual generation, automatic report drafting, user conversations, and so on.

1) Testing Objectives

To evaluate whether the model can generate textual content in high quality based on the input information to meet the practical application requirements.

2) Testing Measurements

Provide diversified generation tasks, including technical document generation, troubleshooting process description, user conversation, and so on.

Verify the accuracy, fluency, information relevance, and diversity of outputs.

3) Evaluation Indexes

Relevance: Whether the generated content is highly relevant to the input information.

$$Relevance = \frac{Relevant\ content\ counts\ generated\ by\ the\ LLM}{Total\ generation\ task\ counts} * 100\%$$

Diversity: Whether the model can generate diversified outputs based on different inputs.

$$Diversity = \frac{Diversified\ outputs\ counts\ generated\ by\ the\ LLM}{Total\ generation\ task\ counts} * 100\%$$

Note: This test evaluates the model's text generation capabilities through diversified generation tasks (technical documents, troubleshooting processes, user conversations, etc.).

5.4 Selection of High-Value Testing Scenarios

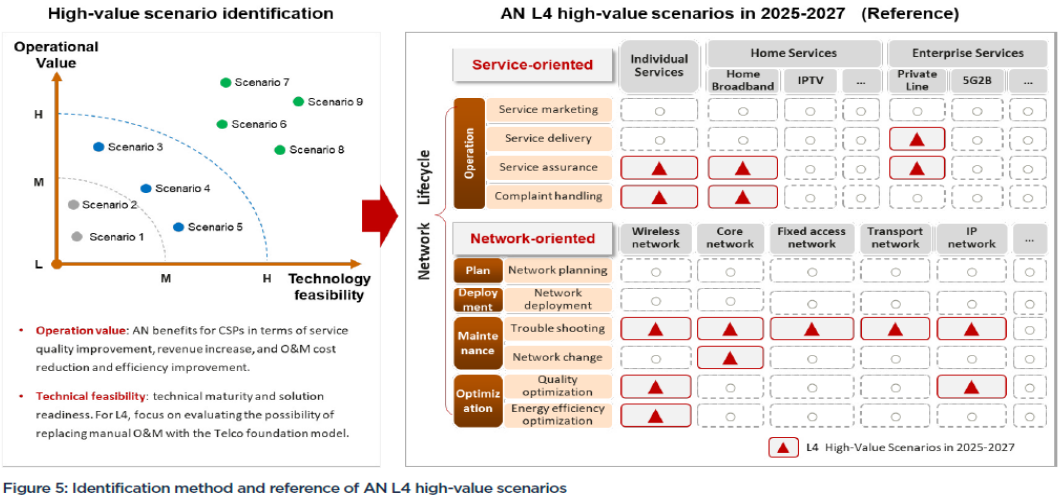


Figure 5: Identification method and reference of AN L4 high-value scenarios

Figure 5-1 Identify AN High-value Scenarios[5]

The application scenarios of autonomous networks run through the full network lifecycle from planning to operation, and each stage highlights its own application scenarios. With partnerships with several communication service providers (CSPs), TMF has evaluated the high value AN scenario for industry references based on the real needs and challenges of network operation. And for telcos, the high-level evolution of autonomous networks is also mainly embodied in the high-value scenarios at the stages at network maintenance, network optimization, and network operations.

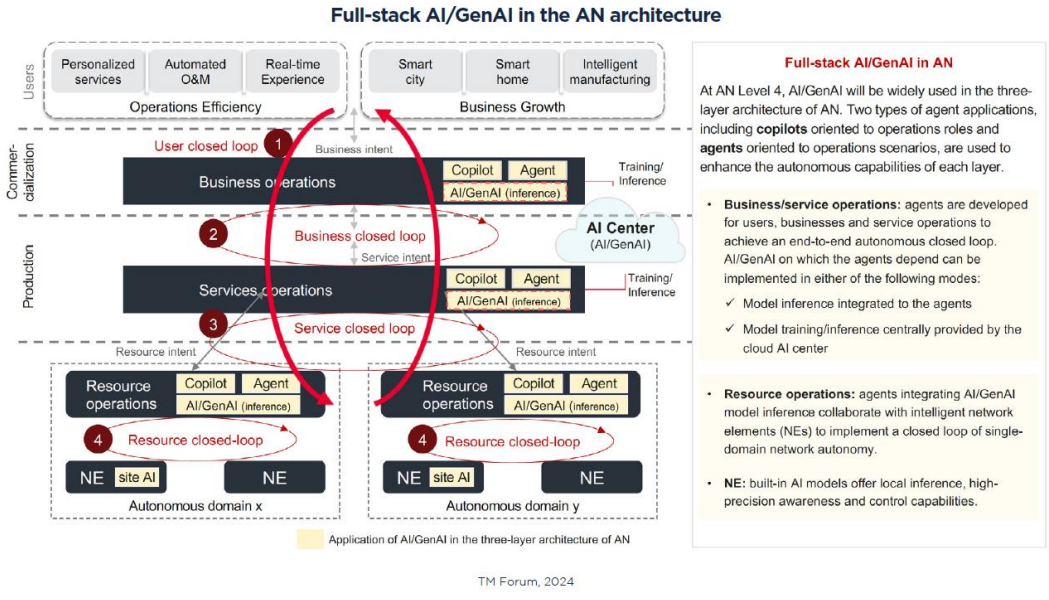


Figure 5-2 Full-stack AI in the AN architecture [1]

According to TMF's AI-enabled AN architecture, LLM empowers autonomous networks through two types of application services: user-oriented Copilot and operation-oriented Agent. LLMs such as DeepSeek can be applied to every level of autonomous network operations among businesses, services, and resources to strengthen closed-loop capabilities.

Combined with AISWare AN Evo and the telcos' business requirements of high-value scenarios in the actual production, eight high-value sub-scenarios are selected from the stages of network maintenance, network optimization, and network operation for DeepSeek’s LLM capability testing.

Table 5-1 Business Scenarios

Network Lifecycle	Tested Scenarios	Test Item for LLM Competence					
		Semantic Parsing	Reasoning	Intent Recognition	Autonomous Planning	Knowledge Retrieval	Text Generation
Network Operation	Intelligent Service Orchestration	●	●			●	
	Network Data Retrieval and Analysis	●		●			
Network Maintenance	Network Topology Generation	●		●			
	Failure Root Cause Analysis		●	●	●	●	
	IP Network Configuration Generation			●		●	●
	Frontline Installation and Maintenance	●		●		●	
	Perception Diagnosis and Analysis	●		●		●	
	Complaint Handling	●		●		●	

6. Typical High-Value Scenario Based Testing Analysis

6.1 Scenario 1: Intelligent Service Orchestration

6.1.1 Scenario Description and Test Instructions

Intelligent service orchestration is a sub-scenario within the service activation process, typically requiring a service orchestration system to implement scenario capabilities. It applies to AI to automate the design of service processes to meet network service requirements, such as network resource activation. LLM is required to be capable of semantic parsing, reasoning, and knowledge retrieval, aiming to accurately recognize business requirements and efficiently build up service processes.

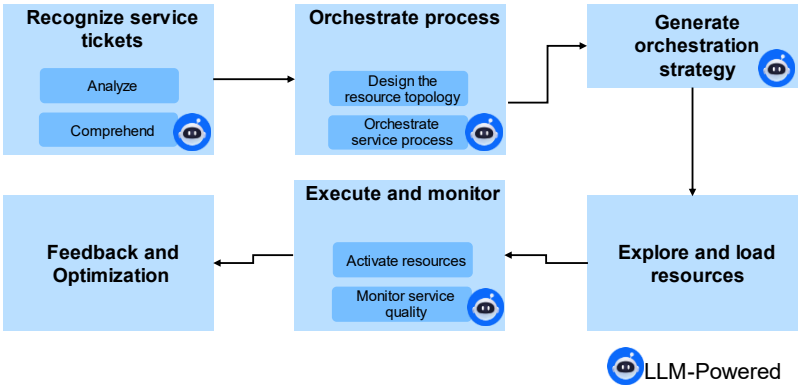


Figure 6-1Design Process of Intelligent Service Orchestration

Testing Steps

1) Design prompt words according to collected information and multi-round interactions.

2) Call DeepSeek R1, DeepSeek V3, and other LLMs respectively, and validate the complete information collection process for each model 800 times.

3) Compare and analyze the capabilities of different models according to the success rate of each model in information collection.

6.1.2 Testing Data Result

Table 6-1 Testing Data of Intelligent Service Orchestration

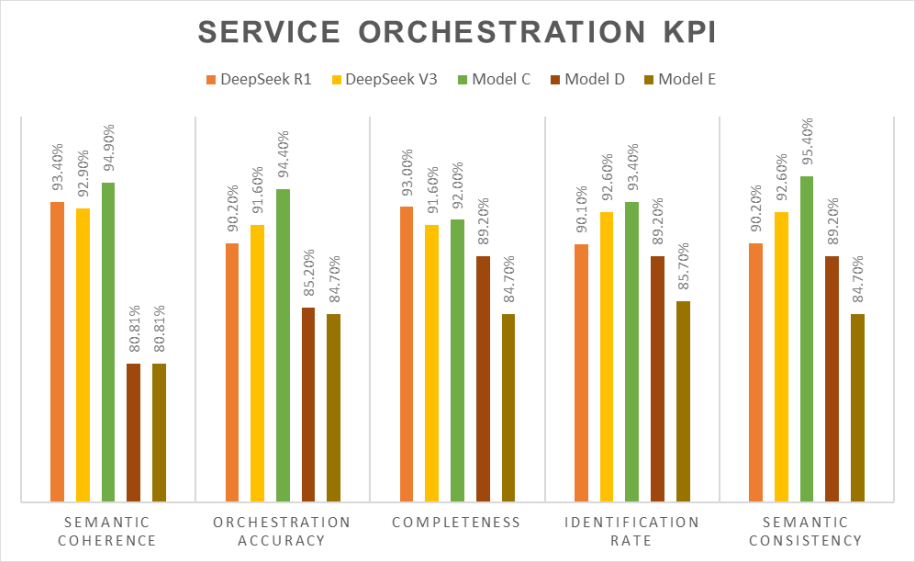
Test item	Data sample	Test data volume
Semantic Parsing of Orchestration Tickets	Question: Design a service process for a private line covering area A with a 100Mbps bandwidth.	800 datasets in AN service

Test item	Data sample	Test data volume
	Expected Output: [%Design a service process for a private line covering area A with a 100Mbps bandwidth on demand, including network topology, device configuration, resource allocation, etc.%]	orchestration scenario
Analytics and Reasoning for Service Processes	Question: Plan a deployment solution for a new service based on existing network resources Expected Output: [%Provide an optimized deployment plan based on existing network resources, including resource scheduling, device configuration, network topology design, etc.%]	800 datasets in AN service orchestration scenario
Knowledge Retrieval in Activation Scenario	Question: Please describe the specific step by step of service activation Expected Output: [%Describe step-by-step of service activation based on standard process, e.g., requirements recognition, equipment installation, functional testing, delivery, etc.%]	800 datasets in AN service orchestration scenario

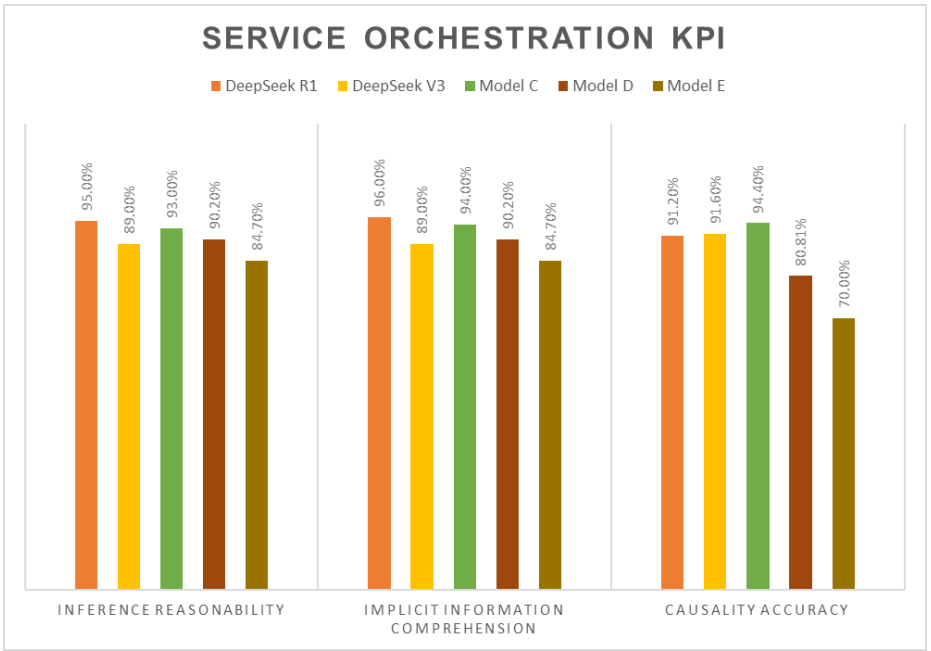
6.1.3 Result Analysis

1) Testing Result Data

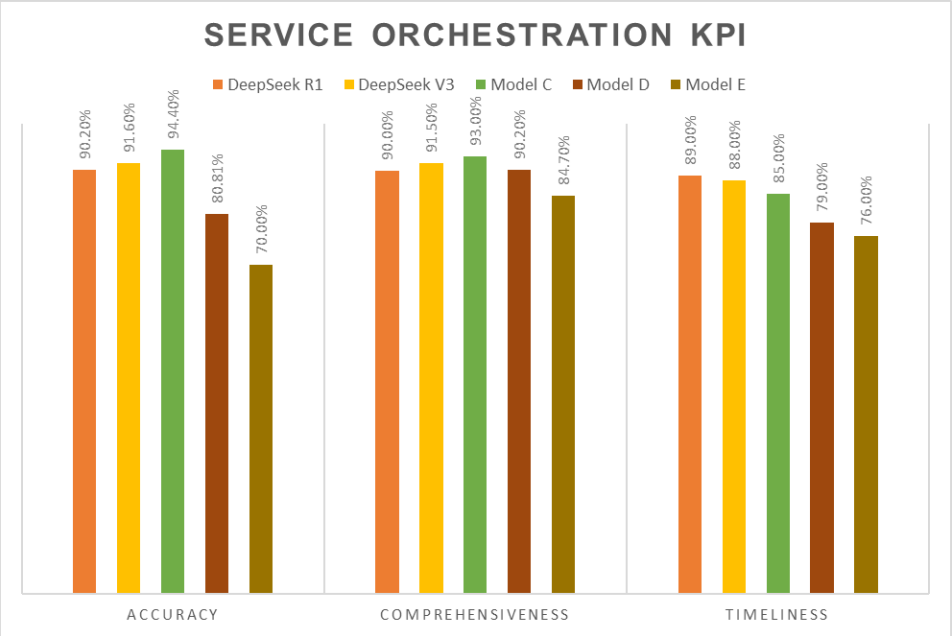
The index definitions and formulas in the testing result for the semantic parsing of intelligent service orchestration are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the reasoning capability of intelligent service orchestration are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the knowledge retrieval of intelligent service orchestration are given in **错误!未找到引用源。**:



2) Testing Data Analytics

Semantic Parsing: DeepSeek R1, DeepSeek V3, and Model C show better under intelligent service orchestration scenarios that reach more than 90% in terms of semantic coherence, accuracy, and completeness and can better understand business requirements and generate reasonable process design. Model D and Model E perform relatively poorly.

Reasoning Capability: DeepSeek R1, DeepSeek V3, and Model C reach more than 90% in terms of reasonableness and completeness in logical deduction and scenario generation and can generate optimized deployment plans based on the existing network resources and providing comprehensive service planning suggestions. In complex service scenarios (e.g., multi-network element co-configuration, cross-domain resource scheduling), these three models show powerful causal reasoning and logic completeness, and the performance results are better. In contrast, Model D and Model E are weaker overall, especially when dealing with multiple constraints or complex tasks, their generated plans have obvious deficiencies in reasonableness and coverage.

Knowledge Retrieval: DeepSeek R1, DeepSeek V3, and Model C show excellent that reach more than 90% in terms of accuracy, comprehensiveness, and timeliness, and can quickly locate key information and answer in detail, especially in standard process descriptions and FAQs, as well as support specific steps of service activation and extraction of related technical details. However, the overall performance of Models D and E is relatively poor, particularly in the case of the standard process descriptions and FAQs,

especially remaining room for improvement in accuracy and comprehensiveness in the retrieval of rare questions or niche technical details.

Performance: The overall performance of the above models can meet the actual production requirements; DeepSeek R1 is relatively slow in generating results after intentional understanding, but it can meet the production needs.

Table 6-2 Testing Result of Service Orchestration

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Intelligent Service Orchestration	Semantic Parsing					
	Reasoning Capability					
	Knowledge Retrieval					

3) Summary

DeepSeek and Model C perform well in intelligent service orchestration, which can accurately fulfill the business requirements and efficiently construct service processes with qualified functionality. Although DeepSeek R1 is relatively slow in generating results, it can still meet production needs.

a) Strengths

DeepSeek R1 and V3 provides high knowledge retrieval accuracy and conforms to standardized process specifications.

b) Deficiencies

DeepSeek-R1 performs poorly under deep-thinking mode and may overthink, which should be further optimized to balance performance and efficiency.

6.2 Scenario 2: Network Data Retrieval and Analysis

6.2.1 Scenario Description and Test Instructions

Network data retrieval and analysis can analyze and predict network services based on network index retrieval to support operation decision-making. This scenario requires LLMs to parse semantics and identify intent to search network data and generate analytical reports efficiently and accurately.

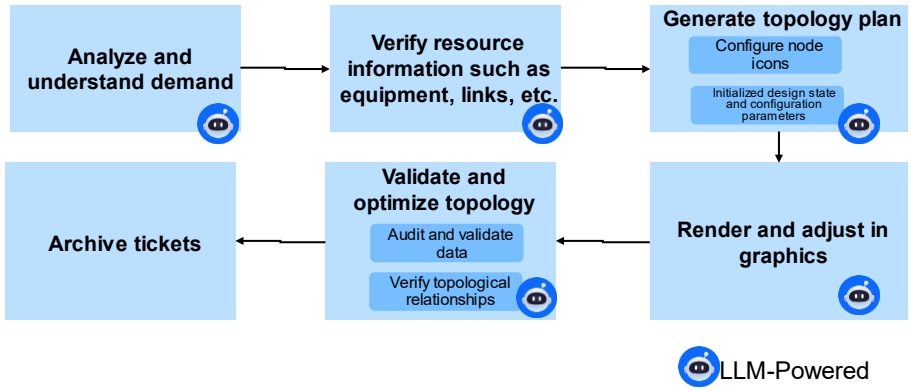


Figure 6-2 Traditional Network Data Retrieval and Analysis Testing Steps

Input network index retrieval commands in natural languages and output retrieval results and analysis reports.

Record evaluation indicators according to the functionality test approach.

Call DeepSeek R1/DeepSeek V3 and other LLMs, respectively, and verify and compare them through standardized test data.

6.2.2 Testing Data Result

Table 6-3 Data of Network Data Retrieval and Analysis

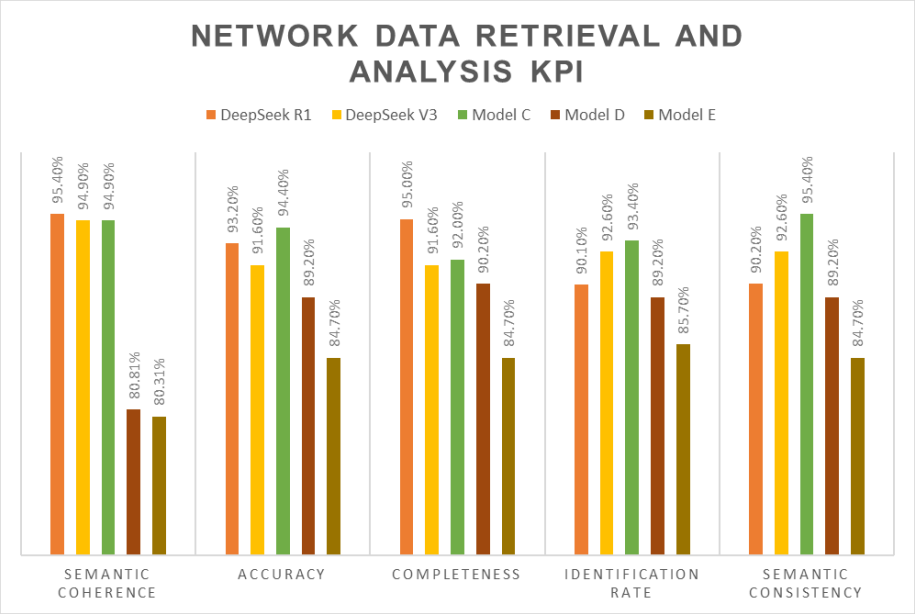
Test Item	Data Sample	Test Data Volume
Semantic Parsing of Network Data Retrieval	Question: Find out the network traffic peaks and the corresponding period in the last week. Expected Output: [%Provide network traffic peaks and corresponding periods, including traffic flow charts, peak periods and other information%]	1200 datasets for network service O&M analytics
Intent Recognition for Report Generation	Question: Analyze the trend of network utilization in the current month Expected Output: [%Provide an analysis of the trend of network utilization during the month, including monthly fluctuations,	1200 datasets for network service

Test Item	Data Sample	Test Data Volume
	major influencing factors, trend forecasts, etc.%]	

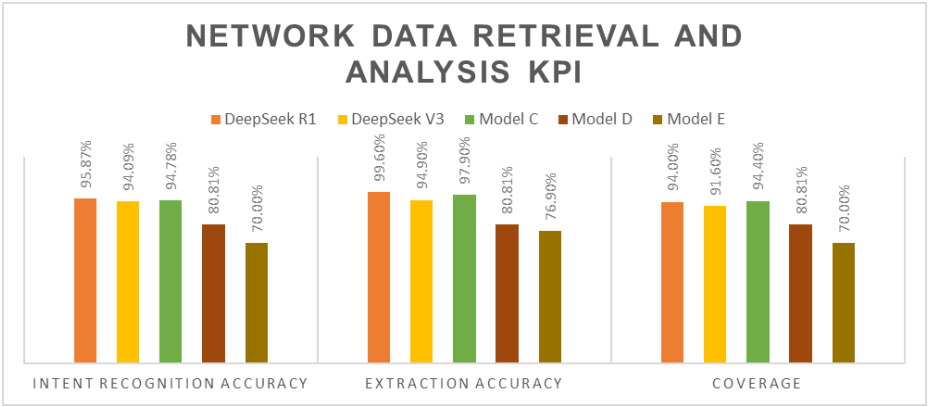
6.2.3 Result Analysis

1) Testing Result Data

The index definitions and formulas in the testing result for the intent recognition of network data retrieval and analysis are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the semantic parsing of network data retrieval and analysis are given in **错误!未找到引用源。**:



2) Testing Data Analytics

Semantic Parsing: DeepSeek R1 and DeepSeek V3 perform well and can accurately understand the retrieval commands and provide the corresponding network data, with more than 90% semantic coherence and accuracy, as well as generating a full-service process design; Model C also performs well, but Models D and E have a slightly lower accuracy when processing complex inquiries.

Intent Recognition: All models reach high accuracy, with DeepSeek R1 and DeepSeek V3 slightly higher than the other models in capturing user intent and generating accurate analytics.

Table 6-2Testing Result of Network Data Retrieval and Analysis

Testing scenario	Required LLM capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Network Data Retrieval and Analysis	Semantic Parsing					
	Intent Recognition					

3) Summary

DeepSeek R1 and DeepSeek V3 performed well under network data retrieval and analysis scenarios, enabling fast and accurate data retrieval and analysis reports generation. Model C also performs good, but Models D and E need to improve their performance when dealing with complex inquiries.

a) Strengths

- DeepSeek R1 and V3 parse the semantics of service scenarios accurately with outstanding process integrity.
- DeepSeek R1 and V3 retrieve knowledge with high accuracy and meet standard process specifications.

b) Deficiencies

- DeepSeek-R1 needs to perform multi-perspective hypothesis deduction and causal chain analysis in deep-thinking mode, resulting in a task understanding and planning delay of 2-5 times that of other models. This mode may be overthinking and needs to be further optimized to balance performance and efficiency.

6.3 Scenario 3: Network Topology Generation

6.3.1 Scenario Description and Test Instructions

Network topology generation is a sub-scenario within network change monitoring, network fault monitoring, and other related scenarios, typically provided by the resource management system. It is to automate the generation of network topology by AI to boost the efficiency of network resource sharing. The LLM is required to equip with intent recognition and semantic parsing under this scenario, which can accurately understand user requirements and quickly generate a reasonable network topology structure.

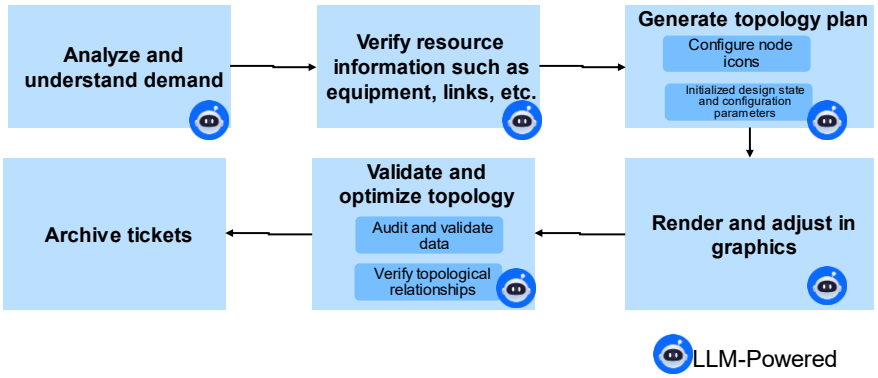


Figure 6-3 Network Topology Generation Process

Testing Steps

1) Input the requirements in natural languages and output the corresponding network topology structure diagram.
2) Record evaluation indicators according to the functionality test approach.
3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare by standardized test data.

6.3.2 Testing Data Result

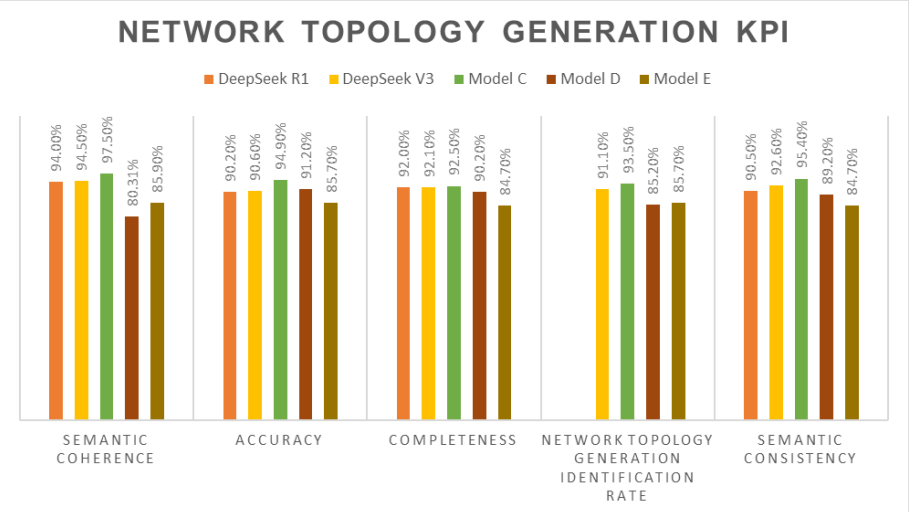
Table 6-5 Data of Network Topology Generation

Test Item	Data Sample	Test Data Volume
Intent Recognition	Question: Generate a topology deployed in a radio access network with 10 network elements. Expected Output: [%Parses the base station name (base station 1, base station 2, etc.), specialty (radio access network), and region name (region name) from the input natural language and outputs a JSON structure%]	600 datasets from network resource operations
Semantic Parsing	Question: Design a network topology for area A with radio and core network elements. Expected Output: [%Parses network element name, specialty (radio access network, core network), region name (region A) from input natural language and outputs JSON structure%]	600 datasets from network resource operations

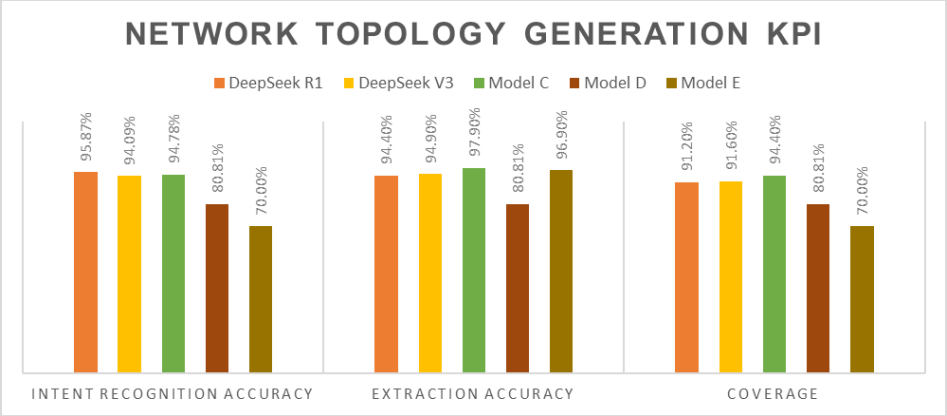
6.3.3 Result Analysis

1) Testing Result Data

The index definitions and formulas in the testing result for the intent recognition of network topology generation are given in 错误!未找到引用源。:



The index definitions and formulas in the testing result for the semantic parsing of network topology generation are given in 错误!未找到引用源。:



2) Testing Data Analytics

Intent Recognition: In the network topology generation scenario, DeepSeek R1 and DeepSeek V3 perform best in intent recognition, accurately understanding user requirements and quickly generating reasonable network topologies. Model C also performs well, but Models D and E fail to accurately recognize user intent in some cases.

Semantic Parsing: All models reached high accuracy, with DeepSeek R1 and DeepSeek V3 slightly higher than the other models, being able to accurately parse out the name of the network element, specialty and area names.

Table 6-6 Testing Result of Network Topology Generation

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Network Topology Generation	Intent Recognition	👍	👍	👍	👎	👎
	Semantic Parsing	👍	👍	👍	👎	👎

3) Summary

Through the overall test, DeepSeek showed a slight decrease in performance without prompt engineering adaptation compared to the fine-tuned Model C. However, this gap is not significant. With proper adaptation, DeepSeek is expected to reach, or even exceed, the level of the Model C after tuning.

a) Strengths

- DeepSeek R1 and V3 generalize across vendor topology specifications with accurate semantic parsing.

b) Deficiencies

- As DeepSeek R1 requires multi-perspective hypothesis deduction and causal chain analysis in deep-thinking mode, its latency for task Parsing and planning is 2-5 times higher than that of other models, and it may suffer from overthinking.

6.4 Scenario 4: Network Failure Root Cause Analysis

6.4.1 Scenario Description and Test Instructions

Network fault root cause analysis is a sub-scenario within network fault monitoring, typically provided by the fault management system. It is to intelligently analyze network failures and pinpoint the root. LLM is required to equip with capabilities of intent recognition, autonomous planning, knowledge retrieval, and reasoning capabilities to quickly diagnose faults and locate root causes.

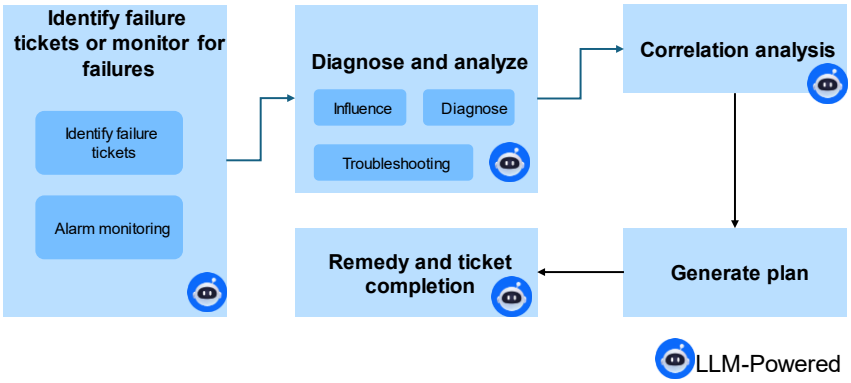


Figure 6-4 Network Failure Resolution

Testing Steps

1) Input network failure description in natural language and output root cause analysis results.

2) Record the evaluation indicators according to the functionality test approach.

3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare them through the above test data.

6.4.2 Testing Data Result

Table 6-7 Testing Data of Network Failure Root Cause Analysis

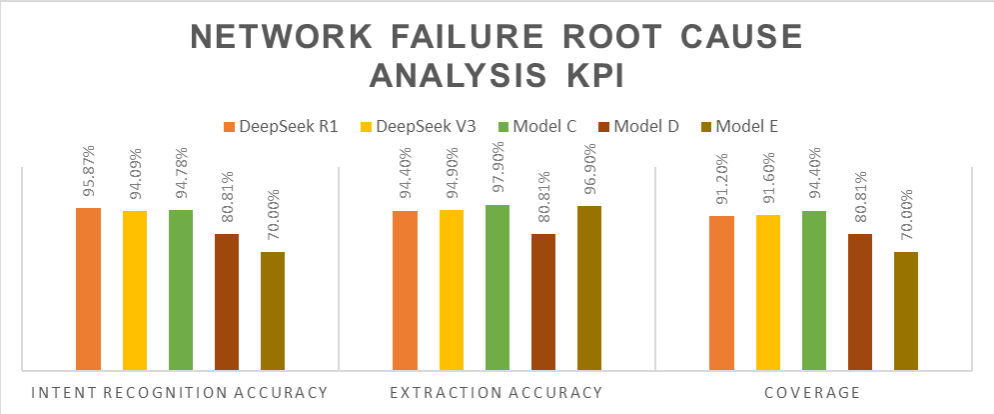
Test Item	Data Sample	Test Data Volume
Intent Recognition	Question: Analyze failed connection of router interface. Expected Output: [%The intent is to analyze the router interface failure to further identify the possible root cause, such as hardware failure, misconfiguration, network traffic congestion, etc.%]	1500 datasets for intelligent network O&M
Autonomous Step Planning	Question: Please plan the steps for troubleshooting. Expected Output: [%Plan detailed troubleshooting steps, including checking router interface configurations, physical connections, network traffic, etc., step-by-step troubleshooting of possible causes, and providing systematic methods%]	1500 datasets for intelligent network O&M
O&M Knowledge retrieval	Question: Find common causes of package loss on switch ports. Expected Output: [%Retrieve common causes of package loss on switch ports based on the network knowledge repository, such as hardware failure, port misconfiguration, poor link quality, network traffic congestion, etc., and output related analytical information%]	1500 datasets for intelligent network O&M
Alarm Localization and Reasoning	Question: Infer the possible failure location based on the alarm. Expected Output: [%Infer the failure location based on the alarm, considering the link status, device status, and network topology, etc., and	1500 datasets for intelligent network O&M

Test Item	Data Sample	Test Data Volume
	gradually troubleshoot the link failures, device breakdown, and finally confirm the location%]	

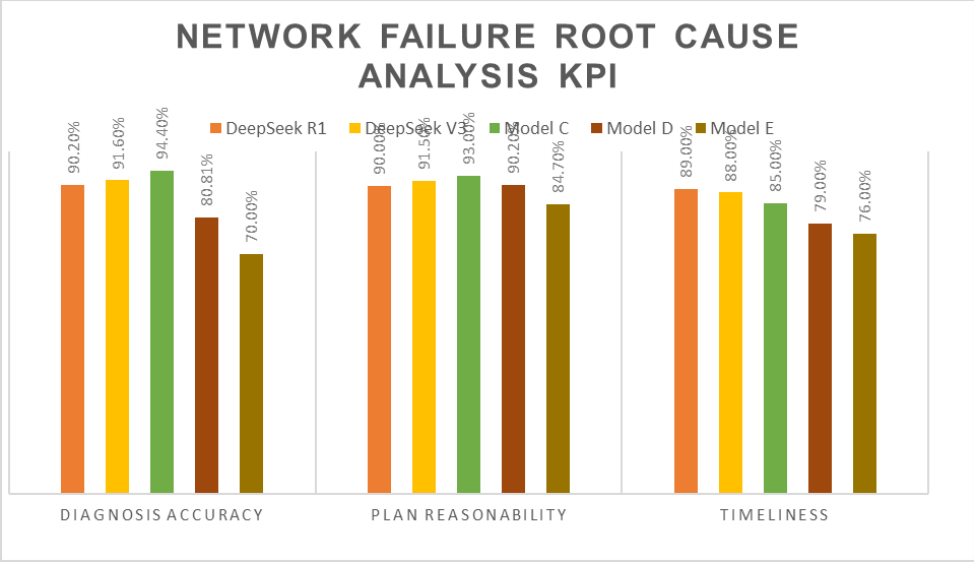
6.4.3 Result Analysis

1) Testing Result Data

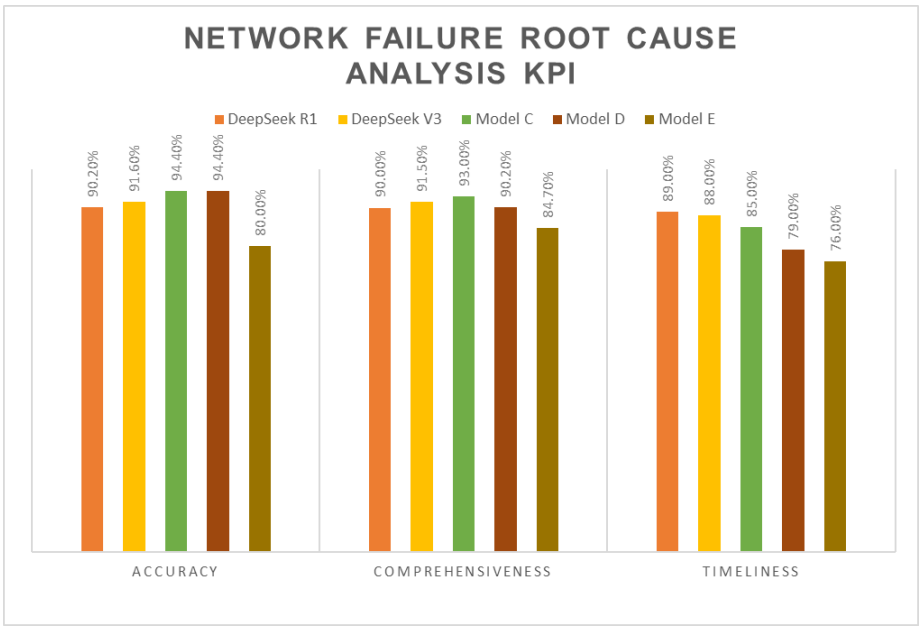
The index definitions and formulas in the testing result for the intent recognition of network failure root cause analysis are given in **错误!未找到引用源。**:



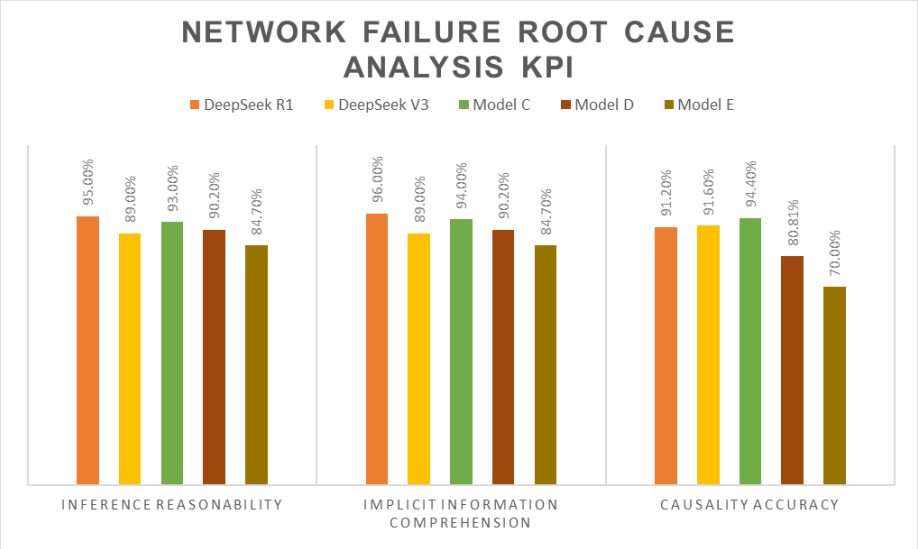
The index definitions and formulas in the testing result for the autonomous planning of network failure root cause analysis are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the knowledge retrieval of network failure root cause analysis are given in 5.3:



The index definitions and formulas in the testing result for the reasoning capability of network failure root cause analysis are given in **错误!未找到引用源。**:



2) Testing Data Analytics

Intent Recognition: DeepSeek R1, DeepSeek V3, and Model C perform the best in accurately understanding user requirements and quickly generating a reasonable network topology. Models D and E fail to accurately recognize user intent in some cases.

Autonomous Planning: DeepSeek R1, DeepSeek V3, Model C, and Model D perform well in generating step-by-step solutions based on existing knowledge and input information. Model E performs slightly less well in terms of reasonableness and timeliness of the generated solutions.

Knowledge Retrieval: DeepSeek R1 and DeepSeek V3 can quickly retrieve relevant failure analysis information from the knowledge base with over 90% accuracy rate. Model C also performs well, while Models D and E need to be improved in terms of accuracy and comprehensiveness.

Reasoning Capability: DeepSeek R1 and DeepSeek V3 perform well in inferring reasonable answers based on common sense and deducing causality or chronological order given the premise, while Model C and Model D can reach similar levels after training, and Model E performs a little less well.

Table 6-8 Testing Result of Generation of Network Topology

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Network Failure Root Cause Analysis	Intent Recognition	👍	👍	👍	👎	👎
	Autonomous Planning	👍	👍	👍	👍	👎

	Knowledge Retrieval					
	Reasoning Capability					

3) Summary

DeepSeek R1 and DeepSeek V3 can accurately analyze network failures and pinpoint the root causes to meet functional requirements in terms of intent recognition, autonomous planning, knowledge retrieval, and reasoning capabilities (except for timeliness), reaching more than 90%.

a) Strengths

- The text quality generated by DeepSeek R1 and V3 is outstanding, and DeepSeek R1 performs excellent in network failure root cause analysis with parameter generalization, especially in terms of intent recognition, autonomous planning, knowledge retrieval, and reasoning capabilities, which are superior to those of the tuned Model C. It can quickly and accurately identify the cause of the failure and provide a solution.

b) Deficiencies

- The deep-thinking mode of DeepSeek R1 leads to the delayed generation of some tasks and requires performance tuning for highly concurrent scenarios.

6.5 Scenario 5: IP Network Configuration Generation

6.5.1 Scenario Description and Test Instructions

IP network configuration generation is to automate the generation of IP configurations and reduce human error. LLM is required to equip with intent recognition, knowledge retrieval, and text generation to quickly generate accurate IP network configuration files.

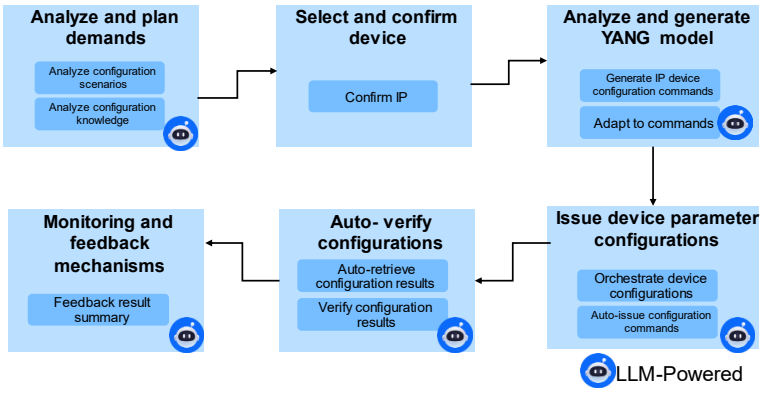


Figure 6-5 Generation process of IP network configurations

Testing Steps

1) Input IP network configuration requirements in natural language and output the generated IP network configuration file.

2) Record the evaluation indicators according to the functionality test approach.

3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare them through the above test data.

6.5.2 Testing Data Result

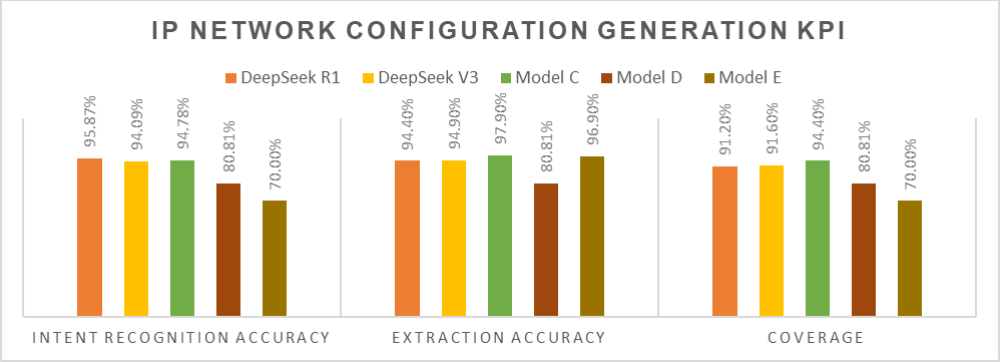
Table 6-9 Testing Data of IP Network Configuration Generation

Test Item	Data Sample	Test Data Volume
Intent Recognition	Question: Generate a subnet configuration with 100 IP addresses. Expected Output: [%Generate a subnet configuration with 100 IP addresses, subnet mask (255.255.255.0), API name (eth0), and other parameters, in correct command format%]	200 datasets for professional IP network O&M
Knowledge Retrieval for Template Generation	Question: Find configuration templates for Router C Expected Output: [%Retrieve configuration templates related to Router C and extract applicable configuration parameters to meet the user's needs%]	200 datasets for professional IP network O&M
Text Generation of Configuration Scripts	Question: Generate a configuration script for Switch D. Expected Output: [%Generate a complete configuration script with multi-steps based on the requirements of Switch D, in correct command format and in accordance with the device specification%]	200 datasets for professional IP network O&M

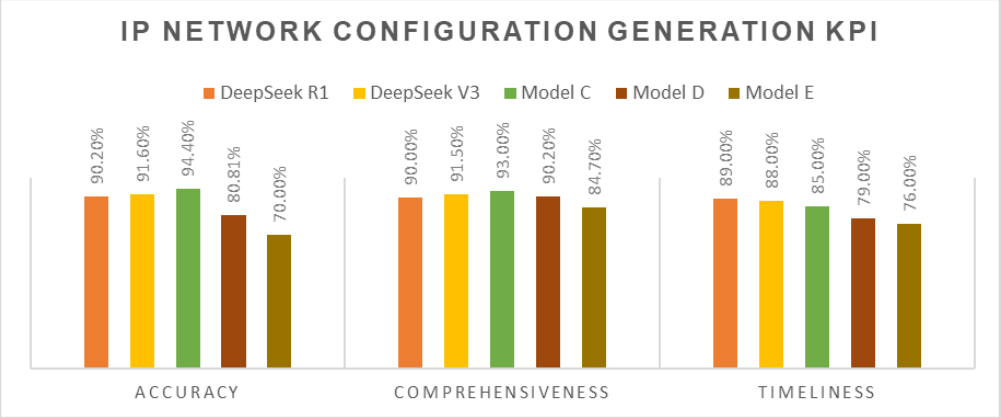
6.5.3 Result Analysis

1) Testing Result Data

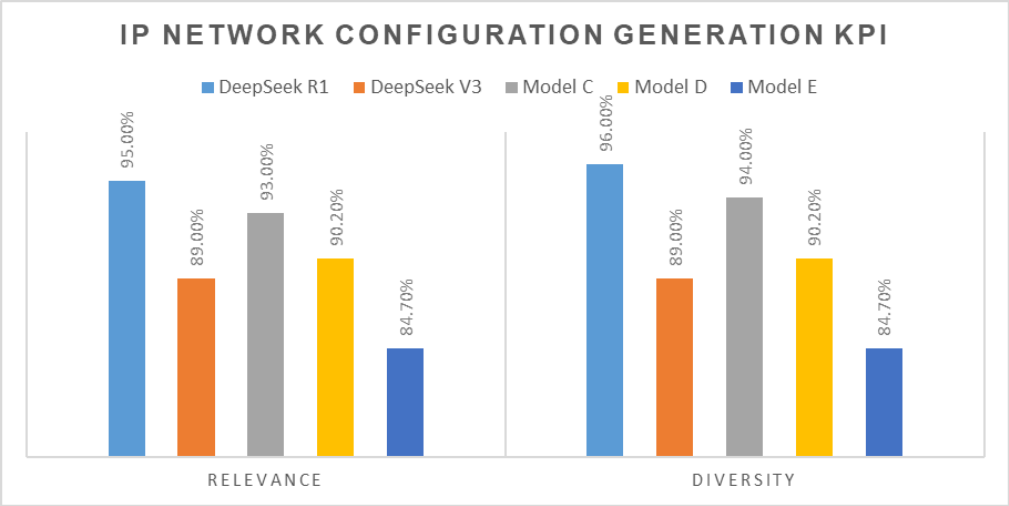
The index definitions and formulas in the testing result for the intent recognition of IP network configuration generation are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the knowledge retrieval of IP network configuration generation are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the text generation of IP network configuration generation are given in **错误!未找到引用源。**:



2) Testing Data Analytics

Intent Recognition: DeepSeek R1 and DeepSeek V3 performed best in accurately understanding the installation and maintenance tasks and providing solutions in terms of intent recognition. Model C also performed well, but Models D and E failed to accurately recognize user intent in some cases.

Semantic Parsing: All models reach a high accuracy rate, with DeepSeek R1 and DeepSeek V3 slightly outperforming the other models in accurately parsing key information.

Knowledge Retrieval: All models perform well and can quickly retrieve relevant installation and maintenance guidance from the knowledge base.

Table 6-10 Testing Results of Frontline Installation and Maintenance Services

Testing scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Frontline Installation and Maintenance Services	Intent Recognition	👍	👍	👍	👎	👎
	Semantic Parsing	👍	👍	👍	👍	👎
	Knowledge Retrieval	👍	👍	👍	👎	👎

3) Summary

DeepSeek R1 and DeepSeek V3 reach more than 91% in terms of semantic coherence and accuracy, and the intention recognition accuracy reaches 95% better than the currently debugged Model C. The knowledge retrieval performance of each model is not very different, mainly relying on the knowledge bases. The text generation of DeepSeek R1 is more effective, with test results exceeding that of the debugged Model C. Comprehensively, the service scenarios can quickly generate accurate IP network configuration files according to the input requirements to satisfy the application scenario.

a) Strengths

- DeepSeek R1 can disassemble complex operations (e.g. fiber optic fusion splicing) clearly, with practical guidance.

b) Deficiencies

- DeepSeek R1 is inadequate in adapting to non-standard device specifications (legacy terminals).

6.6 Scenario 6: Frontline Installation and Maintenance Services

6.6.1 Scenario Description and Test Instructions

In the scenario of handling network complaints in frontline installation and maintenance services, an important sub-scenario is typically handled by the customer complaint support system. But in the advance of high-level autonomous networks, this scenario uses agent methods to enhance on-site network installation and maintenance. LLM is required to equip with intent recognition, semantic parsing, and knowledge retrieval to quickly respond to the installation and maintenance tasks and provide solutions.

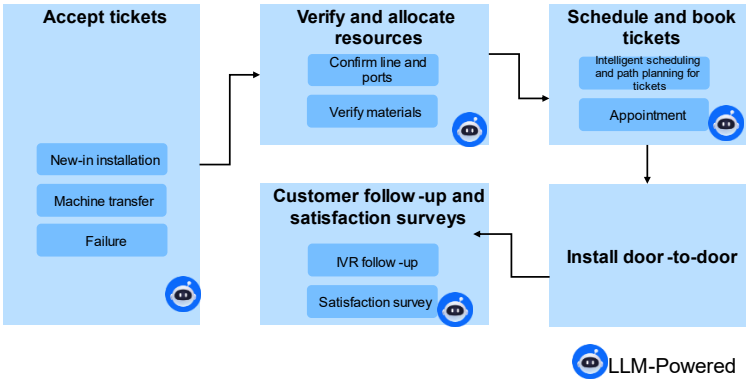


Figure 6-6Service process of frontline installation and maintenance

Testing Steps

1) Input installation and maintenance tasks in natural language and output operation guidance and solutions.

2) Record the evaluation indicators according to the functionality test approach.

3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare them through the above test data.

6.6.2 Testing Data Result

Table 6-11 Testing Data of Frontline Installation and Maintenance Service

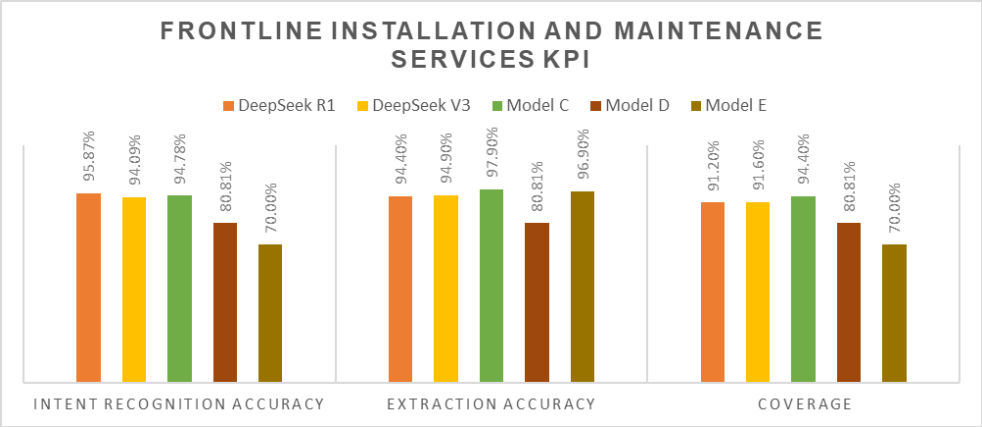
Test Item	Data Sample	Test Data Volume
Intent Recognition	Question: Track the utilization of ONU CPU and ONU memory. Expected Output: Extract keywords, such as “start time”, “indicator”, etc., and organize the information into JSON format according to the prompts in correct fields	1200 datasets for frontline installation and maintenance of home network

Test Item	Data Sample	Test Data Volume
Semantic Parsing	Question: What are the steps to check utilization of ONU CPU and ONU memory? Expected Output: Be able to understand user questions and answer as required.	1200 datasets for frontline installation and maintenance of home network
Knowledge Retrieval	Question: What are the common connections to home internet? Expected Output: Be able to retrieve and quickly generate an answer from the knowledge base based on the question understanding. E.g., common connections including FTTH DSL Cable Modem Mobile bandwidth (e.g. 4G/5G).	1200 datasets for frontline installation and maintenance of home network

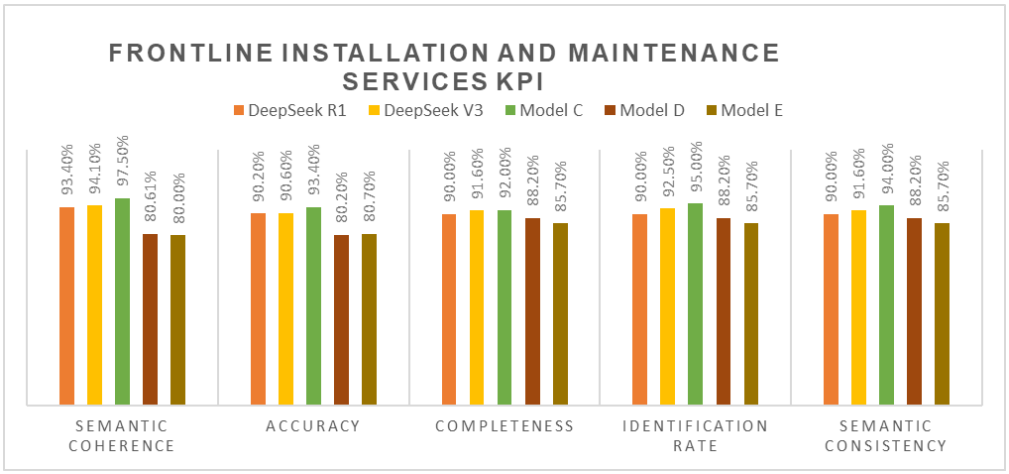
6.6.3 Result Analysis

1) Testing Result Data

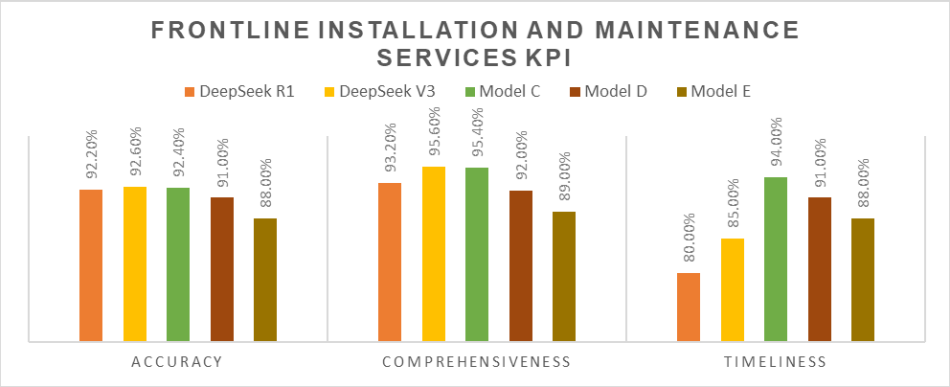
The index definitions and formulas in the testing result for the intent recognition of frontline installation and maintenance service are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the semantic parsing of frontline installation and maintenance service are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the knowledge retrieval of frontline installation and maintenance service are given in **错误!未找到引用源。**:



2) Testing Data Analytic

Intent Recognition: DeepSeek R1 and DeepSeek V3 reach more than 94% in terms of semantic coherence, accuracy, and so on, and some of the indicators are close to or better than those currently tuned in terms of the best performance, able to accurately recognize user intent and provide reasonable solutions. Model C also performed well in knowledge retrieval, but there was not much difference between the models, as Model D and Model E fail to accurately recognize user intent in some cases.

Semantic Parsing: All models reach a high accuracy rate, DeepSeek R1 and DeepSeek V3 are slightly higher than the other models and are able to accurately parse out key information.

Knowledge Retrieval: DeepSeek R1, DeepSeek V3, Model C and Model D all perform well and can quickly retrieve relevant network optimization suggestions from the knowledge base. Model E is less effective.

Table 6-32 Testing Result of Frontline Installation and Maintenance Service

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Frontline Installation and Maintenance Service	Intent Recognition	👍	👍	👍	👎	👎
	Semantic Parsing	👍	👍	👍	👎	👎
	Knowledge Retrieval	👍	👍	👍	👍	👎

3) Summary

DeepSeek R1, DeepSeek V3, and Model C reach better performance in the frontline installation and maintenance service scenarios, Model D can reach better results in knowledge retrieval, and Model E fails to perform well.

a) Strengths

- DeepSeek R1 is highly capable of processing complex tasks, with clear disassembly and practical guidance.

b) Deficiencies

- All models are inadequate in adapting to non-standard device specifications (legacy terminals).

6.7 Scenario 7: Perception Diagnosis and Analysis

6.7.1 Scenario Description and Test Instructions

Perceptual diagnostic analysis is an important sub-scenario in network complaint handling and network performance optimization and is generally provided by the user experience management System for scenario implementation capabilities. Perception diagnosis and analysis is to proactively monitor network quality, prevent and solve potential errors in user experience. LLM is required to equip with the capabilities of intent recognition, semantic parsing, and knowledge retrieval.

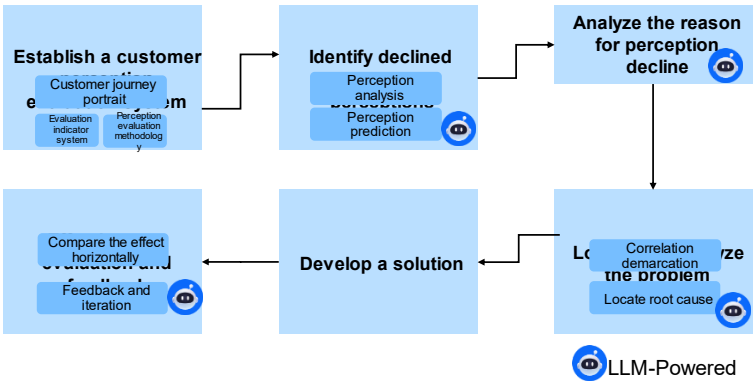


Figure 6-7 Process of Perception Diagnosis and Analysis

Testing Steps

1) Input network quality monitoring in natural language and output monitoring results as well as diagnostic reports.

2) Record the evaluation indicators according to the functionality test approach.

3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare them through the above test data.

6.7.2 Testing Data Result

Table 6-4 Testing Data of Perception Diagnosis and Analysis

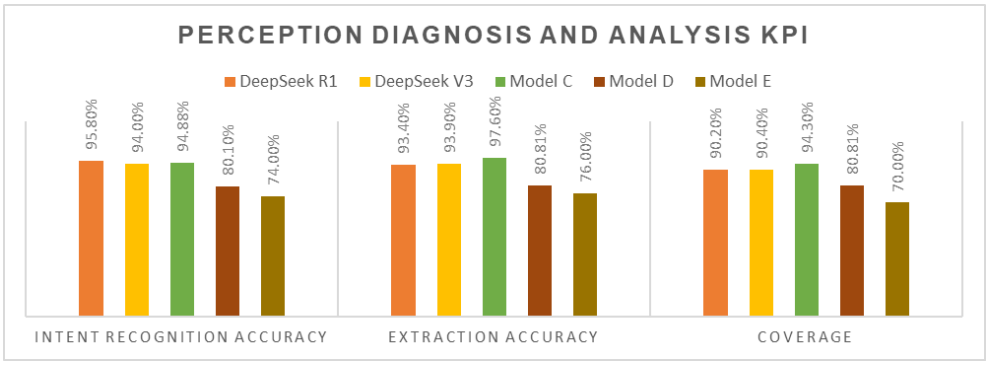
Test Item	Data Sample	Test Data Volume
Intent Recognition	Question: Check the video download rate of user 13075649799 from 3/12/2023 to 3/18/2023. Expected Output: Extract keyword information, such as “time”, “indicator”, etc., and organize the information into JSON format according to the prompts in correct fields.	1000 datasets for network user perception promotion

Test Item	Data Sample	Test Data Volume
Semantic Parsing	Question: What are the steps in the function module Potential Unsatisfied Users Mining? Expected Output: Be able to understand user questions and answer as required.	1000 datasets for network user perception promotion
Knowledge Retrieval	Question: Please describe the user perception portrait. Expected Output: It can retrieve and quickly generate answers from the knowledge base based on an understanding of the question. For example: User perception portrait can evaluate the perception score, perception details, and understand the poor quality events that cause user perception, including user's basic information, service detail, experience dashboard, experience radar, experience indicator details, quality difference list, indicator trend and other modules. This function module is mainly applied to user perception diagnosis, complaint handling analysis, as well as other scenarios.	1000 datasets for network user perception promotion

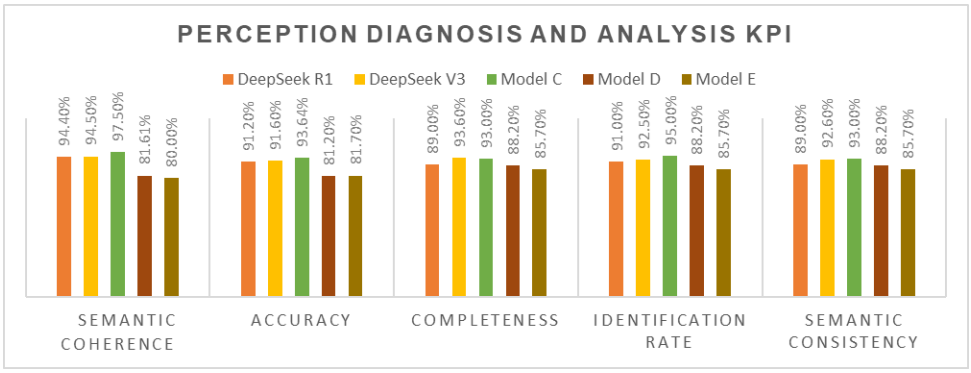
6.7.3 Result Analysis

1) Testing Result Data

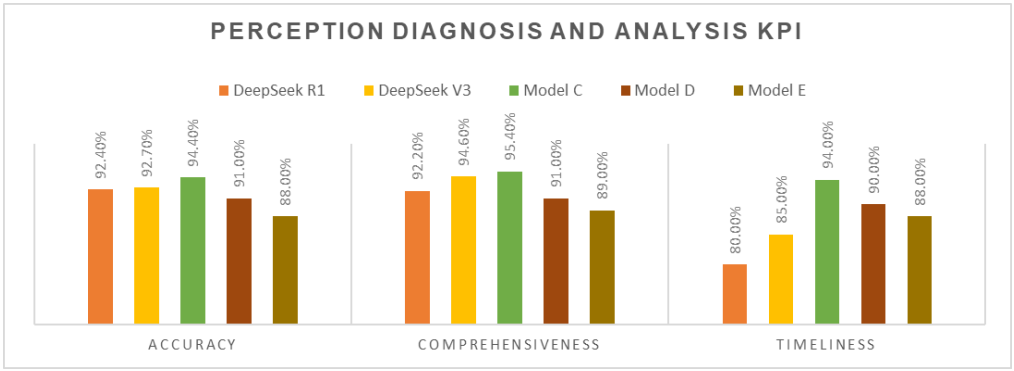
The index definitions and formulas in the testing result for the intent recognition of perception diagnosis and analysis are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the semantic parsing of perception diagnosis and analysis are given in **错误!未找到引用源。**:



The index definitions and formulas in the testing result for the knowledge retrieval of perception diagnosis and analysis are given in **错误!未找到引用源。**:



2) Testing Data Analytics

Intent Recognition: DeepSeek R1 and DeepSeek V3 perform best in accurately recognizing user intent and providing reasonable solutions. Model C also performs well, but Models D and E fail in some cases.

Semantic Parsing: All models reach high accuracy, with DeepSeek R1 and DeepSeek V3 slightly outperforming the other models in accurately parsing key information.

Knowledge Retrieval: All models perform well and can quickly retrieve relevant network optimization suggestions from the knowledge base, while DeepSeek R1 performs poorly in terms of timeliness.

Table 6-54 Testing Result of Perception Diagnosis and Analysis

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Perception Diagnosis and Analysis	Intent Recognition					
	Semantic Parsing					
	Knowledge Retrieval					

3) Summary

In the perception diagnosis and analysis scenario, DeepSeek R1, DeepSeek V3, and Model C have high accuracy in intent recognition and semantic parsing and can accurately parse out key information. Model D performs better only in knowledge retrieval.

a) Strengths

- DeepSeek R1 requires further optimization to improve accuracy and comprehensiveness when dealing with complex network issues, such as multi-domain co-optimization.

b) Deficiencies

- DeepSeek R1 requires real-time optimization, and response speed needs to be improved in high-load or complex retrieval scenarios. DeepSeek R1 requires optimization in terms of real-time performance, as the response speed needs to be improved under high load or in complex query scenarios.

6.8 Scenario 8: RAN Complaint Handling

6.8.1 Scenario Description and Test Instructions

Radio complaint handling is an important sub-scenario in network performance optimization, typically implemented by the network optimization system. In this scenario, the causes of radio network errors reported by users are analyzed, on-site testing is conducted, and optimizations are made through correlation delimitation and quality difference positioning. This process enables quick handling of complaints, improving network quality and enhancing user satisfaction. By applying LLM to identify intentions and reasons, network-related complaints can be categorized according to existing rules.

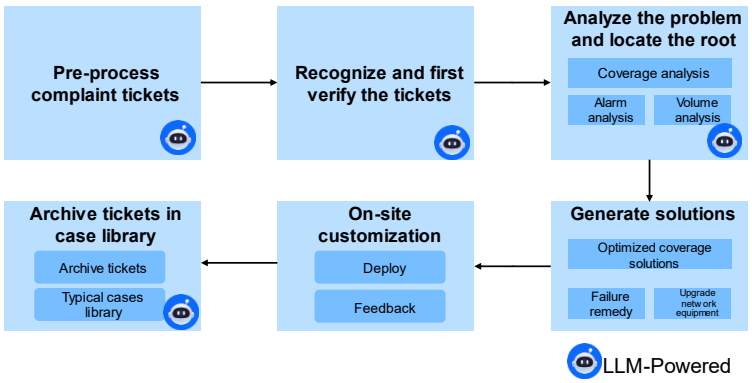


Figure 6-8 Handling Process of Network Complaint

Testing Steps

1) Input in natural language and output question categories.

2) Record the evaluation indicators according to the functionality test approach.

3) Call DeepSeek R1/DeepSeek V3 and other LLMs respectively and validate and compare them through the above test data.

6.8.2 Testing Data Result

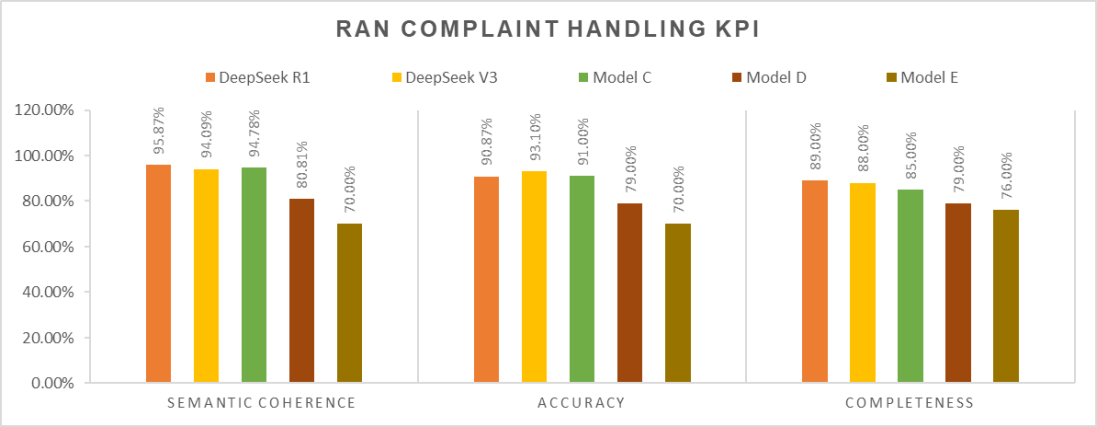
Table 6-65 Testing Data of Complaint Handling

Test item	Data sample	Test data volume
Semantic parsing	Question: Track interference on 394198-71, 2546248-61 on the previous day. Expected Output: [%Calculate from the current date and analyze the radio signal interference situation of the cell on the previous day%]	1000 datasets for complaint handling service
Intent recognition	Question: Users complain about poor Internet signal. Expected Output: [%Internet complaints related data retrieval and analysis%]	1000 datasets for complaint handling service

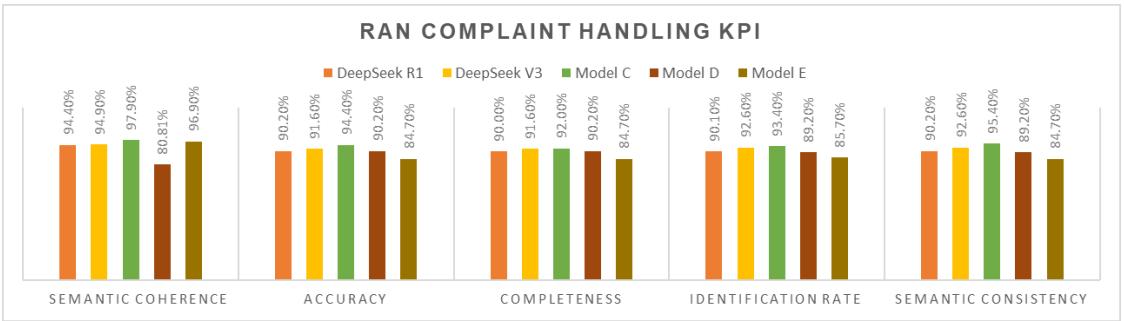
6.8.3 Result Analysis

1) Testing Result Data

The index definitions and formulas in the testing result for the intent recognition of complaint handling are given in 错误!未找到引用源。:



The index definitions and formulas in the testing result for the semantic parsing of complaint handling are given in 错误!未找到引用源。:



2) Testing Data Analytics

Intent Recognition: DeepSeek R1 and DeepSeek V3 perform best in accurately recognizing the intent of user complaints and providing reasonable solutions. Model C also performs well, but Models D and E fails in some cases.

Semantic Parsing: All models reach high accuracy, with DeepSeek R1 and DeepSeek V3 slightly outperforming the other models in accurately parsing key information.

Table 6-76 Testing Result of Complaint Handling

Testing Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	Model C	Model D	Model E
Complaint Handling	Intent Recognition	👍	👍	👍	👎	👎
	Semantic Parsing	👍	👍	👍	👎	👎
	Knowledge Retrieval	👍	👍	👍	👍	👎

3) Summary

DeepSeek R1, DeepSeek V3, and Model C show high accuracy in intent recognition and semantic parsing to accurately capture user intent and parse out key information. DeepSeek R1, DeepSeek V3, Model C, and Model D are quick and responsive in knowledge retrieval and can provide users with timely network optimization recommendations.

a) Strengths

- DeepSeek R1 and V3 are excellent at complaint handling in accurately identifying the user's intent and providing reasonable solutions.

b) Deficiencies

- Further optimization may be required to improve accuracy and comprehensiveness when dealing with rare questions.
- When dealing with complex complaints, such as complaints relating to multiple network devices or multiple outage points, there is still room for improvement in response speed and efficiency.

7. Testing Result Analysis

This test aims to evaluate the technical performance and application potential of DeepSeek in empowering autonomous networks, covering several capabilities applied to autonomous networks, such as intent recognition, semantic parsing, autonomous planning, and reasoning. By accumulating a test dataset of typical high-value scenarios of autonomous networks, DeepSeek and several mainstream domestic LLMs were compared and tested to validate the applicability and efficacy of DeepSeek in autonomous networks.

Table 7-1 Capability Competition among LLMs in Autonomous Network Scenarios

Network Lifecycle	Tested Scenario	Required LLM Capability	DeepSeek R1	DeepSeek V3	C	D	E
Network Operation	Intelligent Service Orchestration	Semantic Parsing	★★★	★★★	★★★	★	★
		Reasoning	★★★	★★	★★★	★★	★
		Knowledge Retrieval	★★★	★★★	★★	★★	★
	Network Data Retrieval and Analysis	Semantic Parsing	★★★	★★★	★★★	★★	★
		Intent Recognition	★★★	★★★	★★★	★★	★
Network Maintenance	Network Topology Generation	Intent Recognition	★★★	★★★	★★★	★★	★★
		Semantic Parsing	★★★	★★★	★★★	★★	★
	Failure Root Cause Analysis	Intent Recognition	★★★	★★★	★★★	★	★★
		Autonomous Planning	★★★	★★★	★★	★★	★
		Knowledge Retrieval	★★	★★	★★★	★★	★
		Reasoning	★★★	★★	★★★	★★	★
	IP Network Configuration Generation	Intent Recognition	★★★	★★★	★★★	★★	★
		Knowledge Retrieval	★★	★★	★★	★★	★
		Text Generation	★★★	★★	★★★	★★★	★

	On-stage Installation and Maintenance	Intent Recognition	★★	★★	★★★	★	★★
		Semantic Parsing	★★★	★★★	★★★	★	★★
		Knowledge Retrieval	★★	★★	★★★	★★	★
Network Optimization	Perception Diagnosis and Analysis	Intent Recognition	★★★	★★★	★★★	★	★
		Semantic Parsing	★★	★★★	★★★	★	★
		Knowledge Retrieval	★★	★★	★★★	★★★	★
	Complaint Handling	Intent Recognition	★★	★★	★★	★	★
		Semantic Parsing	★★★	★★★	★★★	★★	★
		Knowledge Retrieval	★★★	★★★	★★★	★★	★★

Rating: ★★★ (Excellent) ★★ (Average) ★ (Poor)

7.1 DeepSeek Strengths for AN

DeepSeek shows significant technical superiority in several high-value scenarios of the autonomous network, especially in semantic parsing, intent recognition, reasoning capability, autonomous planning, knowledge retrieval and text generation. The following is the perspective of professional empowerment for autonomous networks, combined with specific scenarios and test data, to illustrate the strengths of the DeepSeek:

- Network Failure Root Cause Analysis: DeepSeek R1 and V3 demonstrate excellent intent recognition, autonomous planning, knowledge retrieval, and reasoning capabilities. For example, when dealing with a disconnected router interface, the model can quickly identify the intent, plan detailed troubleshooting steps, and retrieve relevant information from the knowledge base to accurately deduce the failure location. The test data shows that DeepSeek R1 and V3 reach over 90% of the intent recognition accuracy and reasoning capability in this scenario for effective support of quick diagnosis and remedy. In contrast, Model C also performs well in this scenario, but its

reasoning and autonomous planning capabilities are slightly lower to DeepSeek R1 and V3 when dealing with complex failures.

- Autonomous Network Configuration Generation: DeepSeek R1 and V3 outperform in intent recognition, knowledge retrieval and text generation. The models can accurately understand user requirements and quickly generate accurate IP network configuration files. For example, when generating a subnet configuration containing 100 IP addresses, the configuration file generated by the model is formatted correctly with integrated parameters to satisfy the real demands. The test results show that DeepSeek R1 and V3 are superior to other models in terms of intent recognition accuracy and text generation quality, which can effectively reduce human errors and improve network deployment efficiency. Model C also performs well in this scenario, but its accuracy in intent recognition and semantic parsing is slightly lower than that of DeepSeek R1 and V3 when dealing with complex tasks, while Model D performs well, but its accuracy and completeness in text generation is slightly lower than that of DeepSeek R1 and V3.

- Frontline Installation and Maintenance Service: DeepSeek R1 and V3 can significantly enhance the efficiency of field network installation and maintenance by intent recognition, semantic parsing, and knowledge retrieval. For example, the model can accurately recognize the user's intent and provide detailed installation and maintenance instructions when dealing with home internet. Test data shows that DeepSeek R1 and V3 both reach over 94% accuracy in intent recognition and semantic parsing to effectively support frontline installation and maintenance personnel in quick response. Model C also performs well, but its accuracy in intent recognition and semantic parsing is slightly lower than DeepSeek R1 and V3 when dealing with complex tasks.

- Network Optimization and User Perception: In the scenarios of perception diagnosis and analysis, as well as complaint handling, DeepSeek R1 and V3 feature powerful intent recognition, semantic parsing, and knowledge retrieval. The models can monitor network quality in real time, accurately diagnose potential errors, and provide reasonable optimization suggestions.

For example, when dealing with the decreased video download rate, the model can quickly identify the user's intent and retrieve relevant optimization suggestions from the knowledge base. Test results show that DeepSeek R1 and V3 can effectively support network optimization and improve customer perception with over 90% accuracy in both intent recognition and semantic parsing.

In summary, in the high-value scenarios of autonomous network, DeepSeek R1 and V3 perform well in the scenarios of network failure monitoring and diagnosis, configuration generation of autonomous network, frontline installation and maintenance services, and network optimization and user perception, and it is recommended to prioritize the adoption of DeepSeek R1 and V3; Model C also performs well in some scenarios, and it can be a complementary option.

7.2 DeepSeek Deficiencies for AN

It is assumed that in autonomous network scenarios, compared with other models, DeepSeek still has some deficiencies in the three aspects of response speed and efficiency, capability of handling uncommon questions, and overthinking.

- Response Speed and Efficiency: When working on complex tasks, DeepSeek R1 and V3 still have room for improvement in terms of response speed and efficiency. For example, in the scenarios of network failure root cause analysis and intelligent service orchestration, the model needs to deduce hypotheses and analyze the causal chain from multiple perspectives under the deep-thinking mode, which results in 2-5 times higher latency of task Parsing and planning than that of other models. This may affect the utility of the model in scenarios with high concurrency or high real-time requirements. In contrast, Model C performs better in terms of responsiveness and efficiency in these scenarios and is recommended to be used in scenarios with high responsiveness requirements.

- Capability of answering Handling Uncommon Questions: There are some limitations in the performance of DeepSeek R1 and V3 when dealing with

rare, specialized network issues. For example, in the scenarios of frontline installation and maintenance service as well as perception diagnosis and analysis, the accuracy and comprehensiveness of the model in knowledge retrieval still need to be improved when dealing with non-standard equipment specifications (e.g., legacy terminals) or rare failure problems. This may affect the model effectiveness in the actual production environment, and it is recommended to supplement the model with other models or methods.

- Overthinking: Under the deep-thinking mode, DeepSeek R1 and V3 may overthink, resulting in delays in understanding and planning for specific network tasks. For example, in the scenarios of intelligent service orchestration, the model may make too many assumptions and analyze the causal chain when generating the design scheme for complex service processes, which affects the generation speed and efficiency. It is necessary to further optimize the thinking logic of the model to balance performance and efficiency.

7.3 DeepSeek Enhancement for AN Evolution

There are several aspects recommended for improvement and optimization in autonomous network applications, including:

- Enhance Fine-tuning of Models in Practical Applications of Autonomous Networks: Fine tunes the model with the actual autonomous network production environment to further improve the model's adaptability.

- Optimize the Prompting Engineering of the Autonomous Network LLM: Enhance the model performance in different scenarios by optimizing the prompt engineering in professional network service scenarios to meet the actual requirements.

- Maximize Model Application Performance: Further explore how to optimize model utilization parameters for complex tasks and uncommon questions in autonomous network scenarios to improve response speed and efficiency. For example, the speed of task Parsing and planning can be improved by adapting the thinking logic of the model to reduce unnecessary hypothesis

deduction and causal chain analysis.

- Upgrade the Professional Knowledge Base of the Autonomous Network: It is recommended to enhance the development and maintenance of the knowledge base in the professional autonomous network scenarios, and to extend the support for uncommon questions and non-standard equipment specifications to meet the actual demands.

- LLM + RAG: It is recommended to use it in combination with the retrieval-augmented generation (RAG) in the practical autonomous network application. It can maximize LLM performance in autonomous networks, especially when working on complex tasks, and augment the decision-making capability of the model by learning from secondary targets, to better adapt to the diversified needs of autonomous networks and dynamically changing environments. For example, in the scenarios of network failure root cause analysis and perceptual diagnosis and analysis, the accuracy of reasoning capability and knowledge retrieval can be improved through RAG.

8. Conclusion

The analysis of this test concludes that it is recommended to employ DeepSeek as a foundation model in full lifecycle management during the advanced autonomous network evolution towards L4-5 by adopting LLMs. Specific recommendations are as follows:

- Network Failure Monitoring: Close the loop of fault handling with DeepSeek's capabilities in intent recognition, autonomous planning, knowledge retrieval, and reasoning to quickly discover, diagnose, and locate network failures, and improve efficiency and accuracy. For scenarios with high-efficiency requirements, Model C can be selected.

- Autonomous Network Configuration Generation: Automatic generation of accurate network configuration files by applying DeepSeek's capabilities in intent recognition, knowledge retrieval, and text generation to reduce human error and improve network deployment efficiency.

- Frontline Installation and Maintenance Services: Respond quickly to

installation and maintenance tasks and provide effective solutions incorporating DeepSeek's capabilities in intent recognition, semantic parsing, and knowledge retrieval to improve on-site network installation and maintenance service. Model C can also achieve better results.

- User Perception and Radio Network Optimization: Monitor network quality proactively by applying DeepSeek's capabilities in intent recognition, semantic parsing and knowledge retrieval to prevent and resolve potential poor user experiences, as well as accurately understand user complaints and provide reasonable solutions to improve user satisfaction.

- In practice, it is suggested to combine RAG for further effect promotion of DeepSeek in autonomous network and provide powerful technical support for the network intelligence transformation.

To summarize, DeepSeek shows superior technical capabilities in high-value scenarios of autonomous networks, especially in semantic parsing, intent recognition, reasoning, and autonomous planning. Through further optimization, it is expected for further upgrades in processing complex tasks and uncommon questions, providing more effective technical backup for autonomous network evolution towards L4-5.

9. References

[1] DeepSeek, "DeepSeek v3: Technical advancements in multimodal reasoning and code synthesis," DeepSeek, Tech. Rep. DS-2024-003, 2024. [Online]. Available: https://deepseek.com/tech-report/v3. [Accessed: July 10, 2024].

[2] DeepSeek, "DeepSeek-R1: Technical report on long-context MoE models," DeepSeek, Tech. Rep. DS-2023-001, 2023. [Online]. Available: https://deepseek.com/tech-report/r1. [Accessed: July 10, 2024].

[3] DeepSeek. API documentation for DeepSeek-Lite model. (2024). [Online]. Available: https://platform.deepseek.com/api-docs. [Accessed: July 10, 2024].

[4] “IG1401_TMForum_AN_Guide_level_4_industry_blueprint_v6.0.0”

Nov.2024 TM Forum

[5] “TM_Forum_Autonomous_networks_Level_4_industry_blueprint” June.2024 TM Forum

[6] “IG1326_AN_Empowering_DT_Evolving_towards_Level_4_v1.0.0” June.2024 TM Forum

[7] “Autonomous Networks Industry Blueprint: An Evolution Guide towards L4” 2024 TM Forum

[8] “TS 28.100 Management and orchestration; Levels of autonomous network” 3GPP

[9] “TS 28.312 Management and orchestration; Intent driven Management services for mobile networks”, 3GPP

[10] “Autonomous networks: empowering digital transformation for the telecoms industry”, 2019 TM Forum

[11] “Hype Cycle for Operations and Automation in the Communications Industry” 2024 Gartner

[12] "Research Report on the Benchmarking System of Large Models (2024)", China Academy of Information and Communications Technology (CAICT)

[13] "Accuracy Evaluation of Large Models in Industrial Applications", China Industrial Internet Research Institute, March 2024

[14] "Application Evaluation Report of Financial Large Models (2024)", Shanghai Artificial Intelligence Laboratory

[15] "White Paper on Frontier Technologies of Autonomous Networks", China Communications Association, February 2025

[16] " Autonomous Networks Standard", ITU-T, June 18, 2024

[17] " Autonomous Networks Level Leap", TM Forum, May 10, 2023

10. Contact Us

AsiaInfo Technologies (China) Limited

Address: AsiaInfo Plaza, Coutyard#10 East, Zhongguancun Software Park Phase II, Xibeiwang East Road, Haidian District, Beijing, P.R.China

Postcode: 100193

Fax: (+86) 010-82166699

Tel: (+86) 010-82166688

Email：5G@asiainfo.com

Web: www.asiainfo.com

Thank you

Customer Value Innovator & Digital Transformation Promoter with Full-Stack Data Intelligence Capabilities